Epigenetics and DNase-Seq

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2018
Anthony Gitter
gitter@biostat.wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Anthony Gitter, Mark Craven, and Colin Dewey
Goals for lecture

Key concepts

• Importance of epigenetic data for understanding transcriptional regulation
• Predicting transcription factor binding sites
• Gaussian process models
Introduction to epigenetics
Defining epigenetics

• Formally: attributes that are “in addition to” genetic sequence or sequence modifications

• Informally: experiments that reveal the context of DNA sequence
  – DNA has multiple states and modifications

\[
\begin{array}{cccccccccccccccc}
G & A & C & T & A & G & T & G & C & G & T & T & T & A & C & T
\end{array}
\]

\[
\begin{array}{cccccccccccccccc}
G & T & G & C & G & T & T & T & A & C & T
\end{array}
\]

vs.

\[
\begin{array}{cccccccccccccccc}
G & T & G & C & G & T & T & T & A & C & T
\end{array}
\]

\[
\begin{array}{cccccccccccccccc}
G & T & G & C & G & T & T & T & A & C & T
\end{array}
\]

inaccessible

Histones
Importance of epigenetics

Better understand

• DNA binding and transcriptional regulation
• Differences between cell and tissue types
• Development and other important processes
• Non-coding genetic variants (next lecture)
PWMs are not enough

- Genome-wide motif scanning is imprecise
- Transcription factors (TFs) bind < 5% of their motif matches
- Same motif matches in all cells and conditions
PWMs are not enough

- DNA looping can bring distant binding sites close to transcription start sites
- Which genes does an enhancer regulate?

Enhancer: DNA binding site for TFs, can be far from affected gene

Promoter: DNA binding site for TFs, close to gene transcription start site
Mapping regulatory elements genome-wide

- Can do much better than motif scanning with additional data
- ChIP-seq measures binding sites for one TF at a time
- Epigenetic data suggests where some TF binds

Shlyueva Nature Reviews Genetics 2014
DNase I hypersensitivity

- Regulatory proteins bind accessible DNA
- DNase I enzyme cuts open chromatin regions that are not protected by nucleosomes

Nucleosome: DNA wrapped around histone proteins

Wang *PLoS ONE* 2012
Histone modifications

- Mark particular regulatory configurations

- H3 (protein) K27 (amino acid) ac (modification)
DNA methylation

- Reversible DNA modification
- Represses gene expression

Gene “switched on”
- Active (open) chromatin
- Unmethylated cytosines (white circles)
- Acetylated histones

Gene “switched off”
- Silent (condensed) chromatin
- Methylated cytosines (red circles)
- Deacetylated histones
3d organization of chromatin

- Algorithms to predict long range enhancer-promoter interactions
- Or measure with chromosome conformation capture (3C, Hi-C, etc.)

Rao Cell 2014
3d organization of chromatin

- Hi-C produces 2d chromatin contact maps
- Learn domains, enhancer-promoter interactions

Rao *Cell* 2014
Large-scale epigenetic maps

- Epigenomes are condition-specific
- Roadmap Epigenomics Consortium and ENCODE surveyed over 100 types of cells and tissues
Genome annotation

• Combinations of epigenetic signals can predict functional state
  – ChromHMM: Hidden Markov model
  – Segway: Dynamic Bayesian network
Genome annotation

- States are more interpretable than raw data

Ernst and Kellis *Nature Methods* 2012
Predicting TF binding with DNase-Seq
DNase I hypersensitive sites

- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome

Wang PLoS ONE 2012
DNase I footprints

- Distribution of mapped reads is informative of open chromatin and specific TF binding sites.

- TF binding prevents DNase cleavage leaving DNase I “footprint”, only consider 5’ end.
DNase I footprints to TF binding predictions

- DNase footprints suggest that *some* TF binds that location

- We want to know *which* TF binds that location

- Two ideas:
  - Search for DNase footprint patterns, then match TF motifs
  - Search for motif matches in genome, then model proximal DNase-Seq reads

We’ll consider this approach
Protein Interaction Quantification (PIQ)

- Sherwood et al. *Nature Biotechnology* 2014

- **Given**: TF motifs and DNase-Seq reads

- **Do**: Predict binding sites of each TF

Rieck and Wright *Nature Biotechnology* 2014
PIQ main idea

• With no TF binding, DNase-Seq reads come from some background distribution

• TF binding changes read density in a TF-specific way
PIQ main idea

- Shape of DNase peak and footprint depend on the TF

Sherwood *Nature Biotechnology* 2014
PIQ features

• We’ll discuss
  – Modeling the DNase-Seq background distribution
  – How TF binding impacts that distribution
  – Priors on TF binding

• We’ll skip
  – Modeling multiple replicates or conditions, cross-experiment and cross-strand effects
  – Expectation propagation
  – TF hierarchy: pioneers, settlers, migrants
Algorithm preview

- Identify candidate binding sites with PWMs
- Build a probabilistic model of the DNase-Seq reads
- Estimate TF binding effects
- Estimate which candidate binding sites are bound
- Predict pioneer, settler, and migrant TFs
DNase-Seq background

• Each replicate is noisy, don’t want to over-interpret this noise
  – Only counting density of 5’ ends of reads

• Manage two competing objectives
  – Smooth some of the noise
  – Don’t destroy base pair resolution signal
Gaussian processes

- Can model and smooth sequential data
- Bayesian approach
- Jupyter notebook demonstration
TF DNase profile

- Adjust the log-read rate by a TF-specific effect at binding sites

\[ \hat{\mu}_l = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \leq W \text{ and } I_m = 1 \\ 0 & \text{otherwise} \end{cases} \]

- DNase log-read rate adjusted for binding of factor \( l \)
- DNase log-read rate at position \( i \) from Gaussian process
- Location of binding site \( m \)
- Whether site \( m \) is bound
- Window size
TF DNase profile

- DNase profiles represented as a vector for each TF

\[ \hat{\mu}_l = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \leq W \text{ and } I_m = 1 \\ 0 & \text{otherwise} \end{cases} \]

Can’t be too far apart
Priors on TF binding

- TF binding event $I_j$ should be more likely when
  - motif score $s_j$ is high
  - DNase counts $c_j$ are high

- Isotonic (monotonic) regression

\[
\log(P(I_j = 1)) = f(s_j) + g(c_j)
\]
Full algorithm

• **Given**: TF motifs and DNase-Seq reads
• **Do**: Predict binding sites of each TF

- Identify candidate binding sites with PWMs
- Fit Gaussian process parameters for background
- Estimate TF binding effects $\beta_{i-j,l}$
- Iterate until parameters converge
  - Estimate Gaussian process posterior with expectation propagation
  - Estimate expectation of which candidate binding sites are bound
  - Update monotonic regression functions for binding priors
TF binding hierarchy

- Pioneer, settler, and migrant TFs

Sherwood *Nature Biotechnology* 2014
Evaluation: confusion matrix

- Compare predictions to actual ground truth (gold standard)

Lever Nature Methods 2016
Evaluation: ChIP-Seq gold standard

![Diagram showing motif occurrence, DNase I cut count, predicted binding of X, and actual binding of X by ChIP-seq.](image)
Evaluation: ROC curve

- Calculate receiver operating characteristic curve (ROC)
- True Positive Rate versus False Positive Rate
- Summarize with area under ROC curve (AUROC)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Includes true negatives
Reason to prefer precision-recall for class imbalanced data
Evaluation: ROC curve

- TPR and FPR are defined for a set of positive predictions
- Need to threshold continuous predictions
- Rank predictions
- ROC curve assesses all thresholds

<table>
<thead>
<tr>
<th>Candidate binding site</th>
<th>( P\text{(bound)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>764</td>
<td>0.99</td>
</tr>
<tr>
<td>47</td>
<td>0.96</td>
</tr>
<tr>
<td>942</td>
<td>0.91</td>
</tr>
<tr>
<td>157</td>
<td>0.87</td>
</tr>
<tr>
<td>79</td>
<td>0.83</td>
</tr>
<tr>
<td>202</td>
<td>0.72</td>
</tr>
<tr>
<td>356</td>
<td>0.66</td>
</tr>
<tr>
<td>679</td>
<td>0.51</td>
</tr>
<tr>
<td>291</td>
<td>0.43</td>
</tr>
<tr>
<td>810</td>
<td>0.40</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Calculate TPR and FPR at all thresholds \( t \)
PIQ ROC curve for mouse Ctcf

- Compare predictions to ChIP-Seq
- Full PIQ model improves upon motifs or DNase alone

Sherwood Nature Biotechnology 2014
PIQ evaluation

• Compare to two standard methods
  – 303 ChIP-Seq experiments in K562 cells
  – Centipede, digital genomic footprinting

• Compare AUROC
  – PIQ has very high AUROC
  – Mean 0.93
  – Corresponds to recovering median of 50% of binding sites
DNase-Seq benchmarking

- PIQ among top methods in large scale DNase benchmarking study
- HMM-based model HINT was top performer
Downside of AUROC for genome-wide evaluations

Almost all methods look equally good when using full ROC curve AUROC close to 1.0

Precision-recall curve or truncated ROC curve differentiate methods

Gusmao Nature Methods 2016
PIQ summary

• Smooth noisy DNase-Seq data without imposing too much structure

• Combine DNase-Seq and motifs to predict condition-specific binding sites

• Supports replicates and multiple related conditions (e.g. time series)